IBM DATA SCIENCE CAPSTONE FINAL PRESENTATION

Laureen Luo

Sept 2021

AGENDA

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

EXECUTIVE SUMMARY

- Summary of Methodologies:
  - Data Collection via API and Web Scraping
  - Data Wrangling and Analysis
  - Folium Maps
  - Model Analysis
- Summary of Results:
  - Data Analysis with Visualization
  - Recommended Model for Predictive Analysis

INTRODUCTION

- Project Background and Context:
  - SpaceX advertises the Falcon9 rocket launches on its website for $62M USD, while other providers cost upwards of $165M USD. Much of the savings is due to SpaceX being able to reuse the first stage. If we can predict if the first stage will land, we can determine the launch cost and can be utilized by alternates for bids against SpaceX.
- Problems/Questions to Answer:
  - Will Falcon9 land successfully?
  - Effect of each relationship between rocket variables
  - Conditions to which Falcon9 will have the greatest probability of landing successfully

METHODOLOGY

- Data Collection
  - SpaceX Rest API
  - Web Scrapping Wikipedia
- Data Wrangling
  - Hot encoding data fields and dropping irrelevant columns
- Exploratory Data Analysis (EDA) using visualization and SQL
  - Scatter/bar graphs
- Interactive Visual Analytics using Folium and Plotly
- Predictive Analysis using classification models
  - Classification models

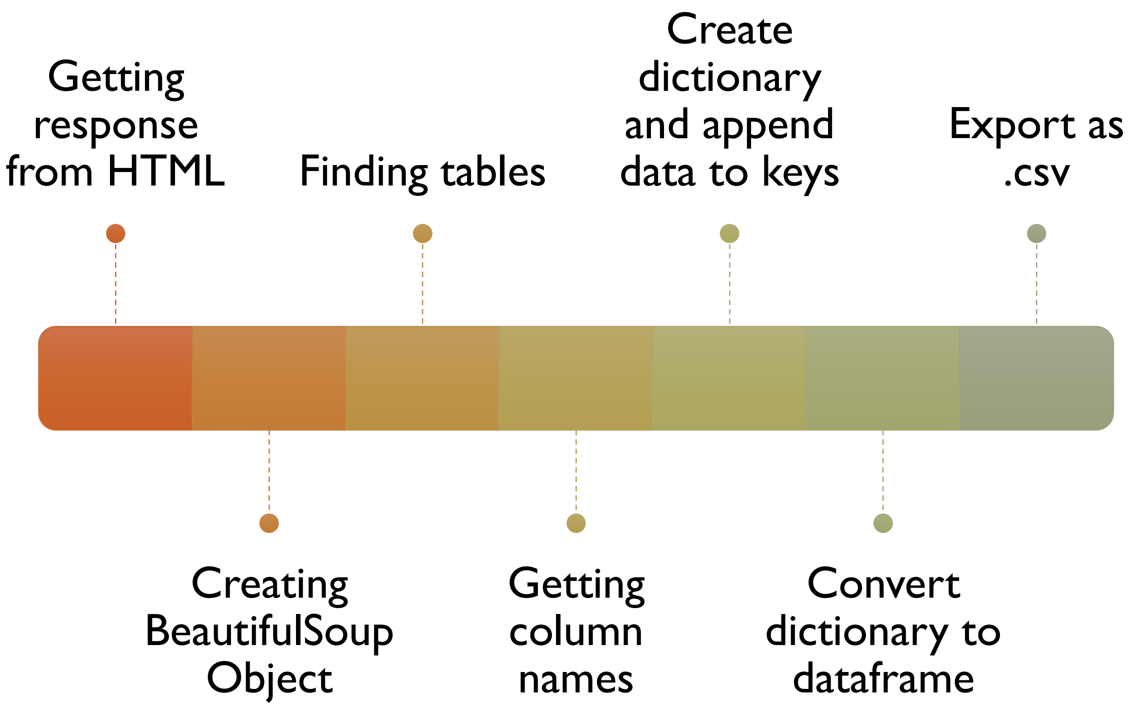
DATA COLLECTION VIA SPACEX API

1. Getting a response from the API
2. Converting to a .json file
3. Clean the data
4. Assign the list to a dictionary and convert to a dataframe
5. Filter dataframe and export as a flat file

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
1	2010-06-04	Falcon 9	6123.547647	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0003
2	2012-05-22	Falcon 9	525.000000	LEO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0005
3	2013-03-01	Falcon 9	677.000000	ISS	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B0007
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	None	1.0	0	B1003
5	2013-12-03	Falcon 9	3170.000000	GTO	CCSFS SLC 40	None None	1	False	False	False	None	1.0	0	B1004

[Github Link]

DATA COLLECTION VIA WEB SCRAPING



Flight No.	Launch site	Payload	Payload mass	Orbit	Customer	Launch outcome	Version Booster	Booster landing	Date	Time
1	CCAFS	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	F9 v1.0B0003.1	Failure	4 June 2010	18:45
2	CCAFS	Dragon	0	LEO	NASA	Success	F9 v1.0B0004.1	Failure	8 December 2010	15:43
3	CCAFS	Dragon	525 kg	LEO	NASA	Success	F9 v1.0B0005.1	No attempt	22 May 2012	07:44
4	CCAFS	SpaceX CRS-1	4,700 kg	LEO	NASA	Success	F9 v1.0B0006.1	No attempt	8 October 2012	00:35
5	CCAFS	SpaceX CRS-2	4,877 kg	LEO	NASA	Success	F9 v1.0B0007.1	No attempt	1 March 2013	15:10

[Github Link]

DATA WRANGLING

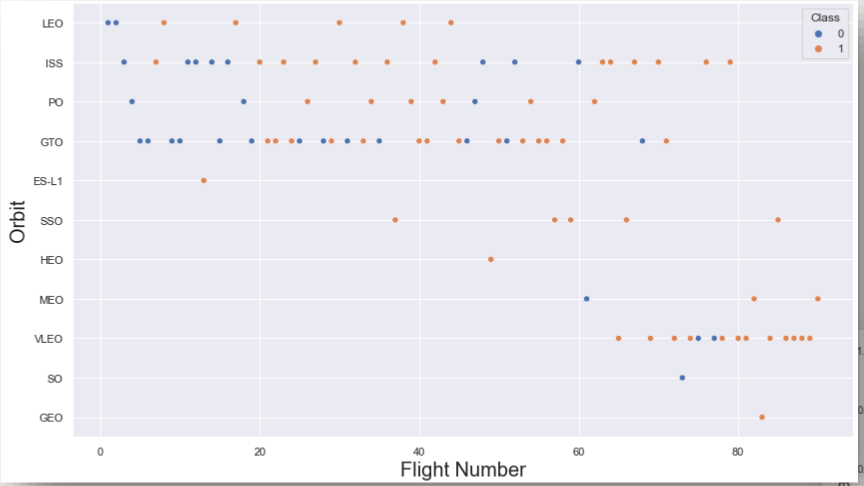
- Calculate number of launches at each site
- Calculate number and occurrence of each orbit
- Calculate number and occurrence of mission outcome per orbit type
- Create landing outcome label from Outcome column
- Create dictionary and append data to keys
- Export as .csv

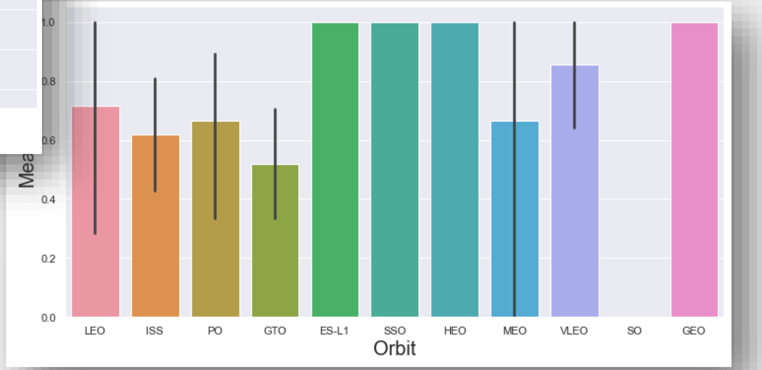
FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block	ReusedCount	Serial
1	2010-06-04	Falcon 9	6104.959412	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0003
2	2012-05-22	Falcon 9	525.000000	LEO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0005
3	2013-03-01	Falcon 9	677.000000	ISS	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B0007
4	2013-09-29	Falcon 9	500.000000	PO	VAFB SLC 4E	False Ocean	1	False	False	False	NaN	1.0	0	B1003
5	2013-12-03	Falcon 9	3170.000000	GTO	CCAFS SLC 40	None None	1	False	False	False	NaN	1.0	0	B1004

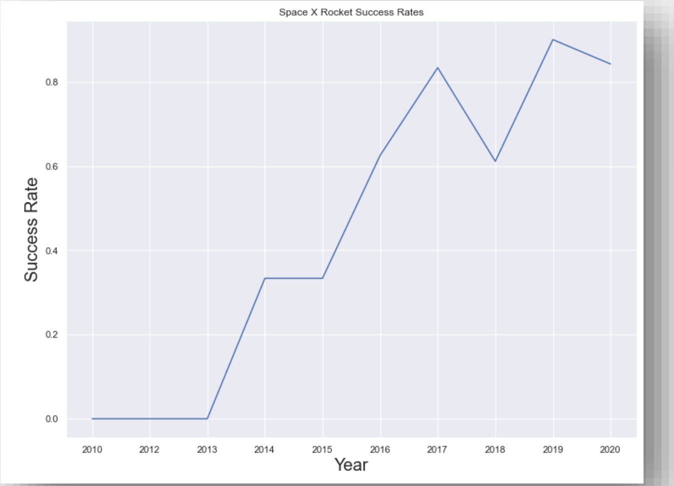
Github Link

EDA WITH DATA VISUALIZATION

- Scatter graph – to determine whether there is a noticeable dependency between the attributes
- Bar graph – to help identify any visual trends or relationships
- Line graph – helps to track the direct relationship and pattern between the data points







Github Link

EDA WITH SQL

- Queries performed:
  - Display the names of the unique launch sites
  - Display 5 records where launch sites begin with the string 'CCA'
  - Display total payload mass carried by boosters launched by NASA
  - Display average payload mass carried by booster version F9
  - List the date where the successful landing outcome in drone ship was achieved
  - List the names of the boosters which have success in ground pad and have a payload mass greater than 4000 and less than 6000
  - List the total number of successful and failed mission outcomes
  - List the names of the booster_versions which have carried the maximum payload mass
  - List the failed landing_outcomes in drone ship, booster versions and launch site names in 2015
  - Rank the count of landing outcomes between 2010 and 2017

Github Link

FOLIUM INTERACTIVE MAP

- Map objects added:

Map Objects	Code	Result
Map marker	Folium.marker(	Added a map marker
Icon marker	Folium.icon(	Added an icon
Circle marker	Folium.circle(	Added a circle
Polyline	Folium.polyline(	Added a line between the points
Marker Cluster marker	MarkerCluster(	Clusters the markers
AntPath	Folium.plugins.antpath(	Added an animated line between the points

PLOTLY DASH

- With Plotly Dash, these objects were added:

Map Objects	Code	Result
Dash	Import dash Import dash_html_components as html	Initiates the data visualization tools
Pandas	Import pandas as pd	Initiates dataframe tools
Plotly	Import plotly.express as px	Plot the graphs with plotly
Dropdown	dcc.Dropdown(	Added dropdown for launch sites
Rangeslider	dcc.RangeSlider(	Added range slider
Pie chart	Px.pie(	Added pie graph
Scatter chart	Px.scatter(	Added scatter graph

PREDICTIVE ANALYSIS (CLASSIFICATION)

- Built the model
  - Load the engineered data into a dataframe
  - Transform and standardize the data using Numpy
  - Split the data into training and test data sets
  - Check how many samples were created and set our parameters/algorithms
  - Fit the datasets into the GridSearchCV objects to train the model
- Evaluating the model
  - Check the accuracy of the model and plot the Confusion Matrix
- Finding the best performing classification model
  - Use the highest accuracy score

FLIGHT NUMBER VS. LAUNCH SITE

- With higher flight numbers (> 30), the success rate increases.

[Github Link]

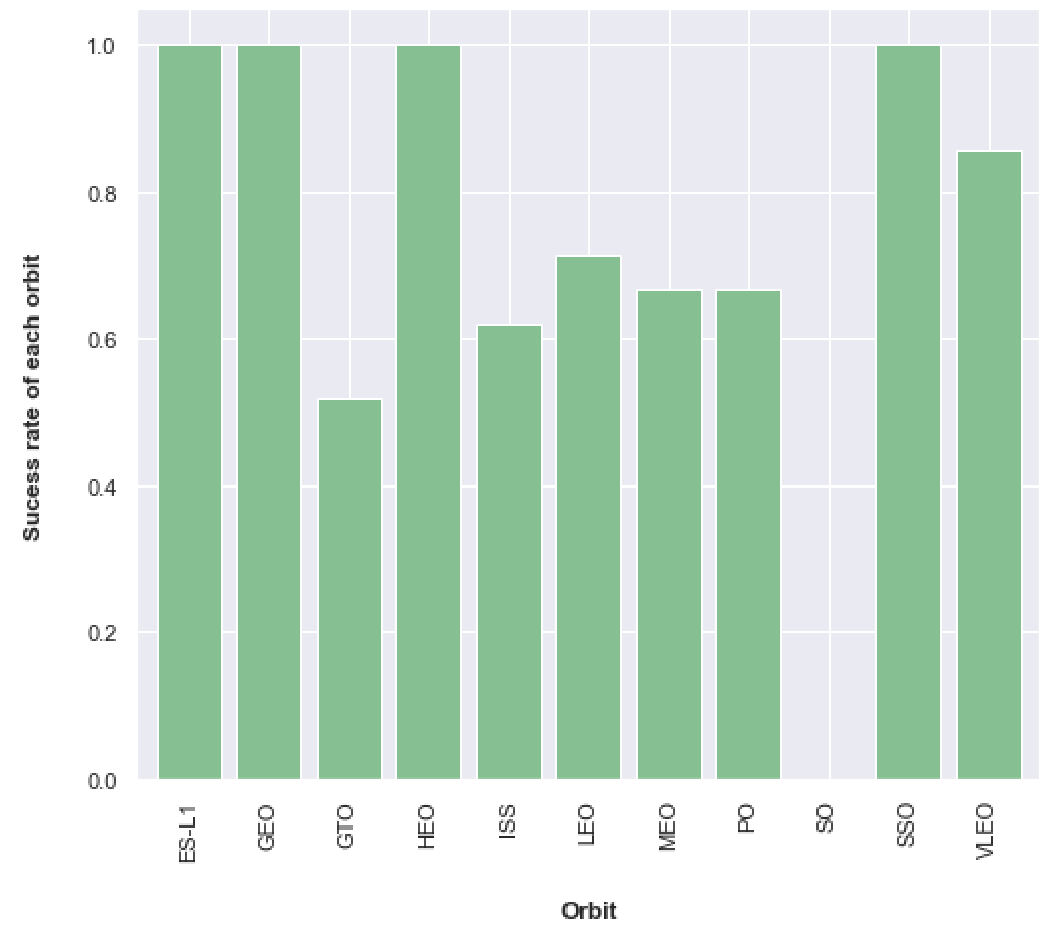
PAYLOAD VS LAUNCH SITE

- With greater payload mass (> 7000 KG), the higher the success rate for the rocket but payload mass and launch site are not directly correlated.

Github Link

SUCCESS RATE VS. ORBIT TYPE

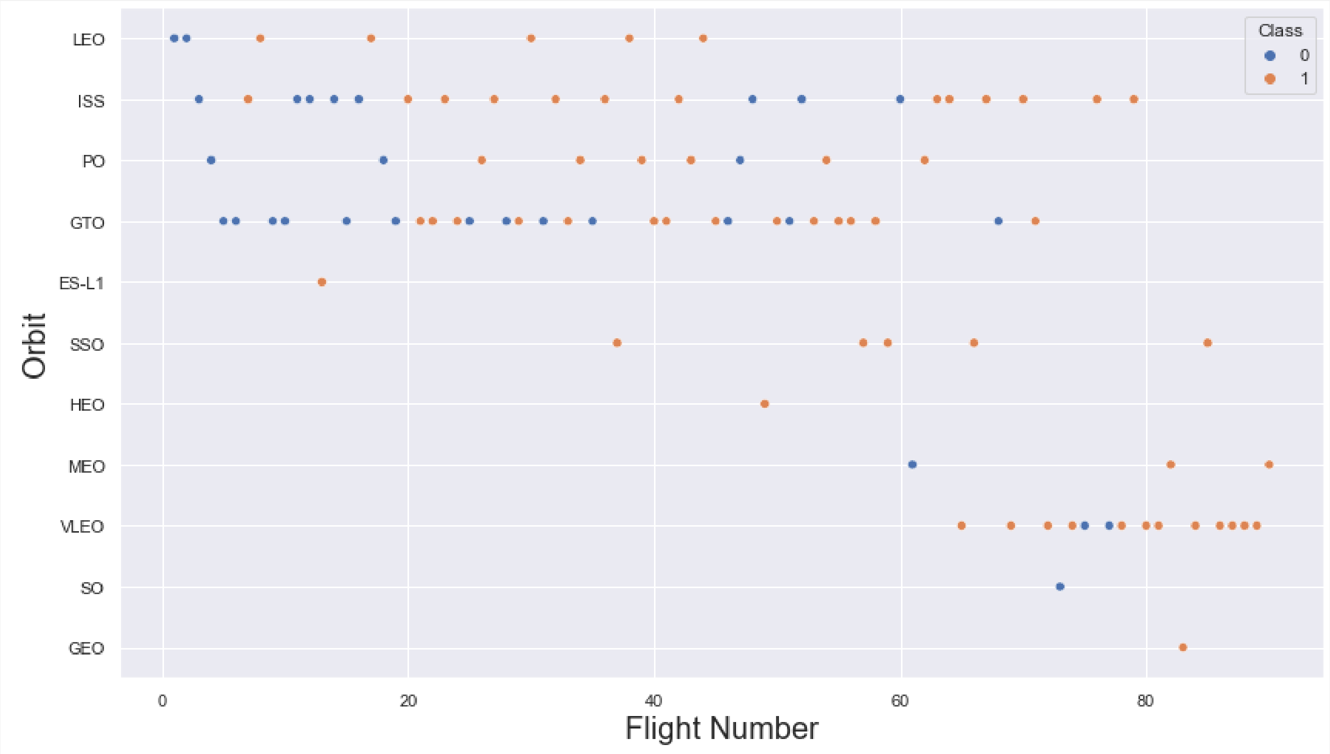
- The ES-L1, GEO, HEO and SSO orbits had the highest success rate.



[Github Link]

FLIGHT NUMBER VS ORBIT TYPE

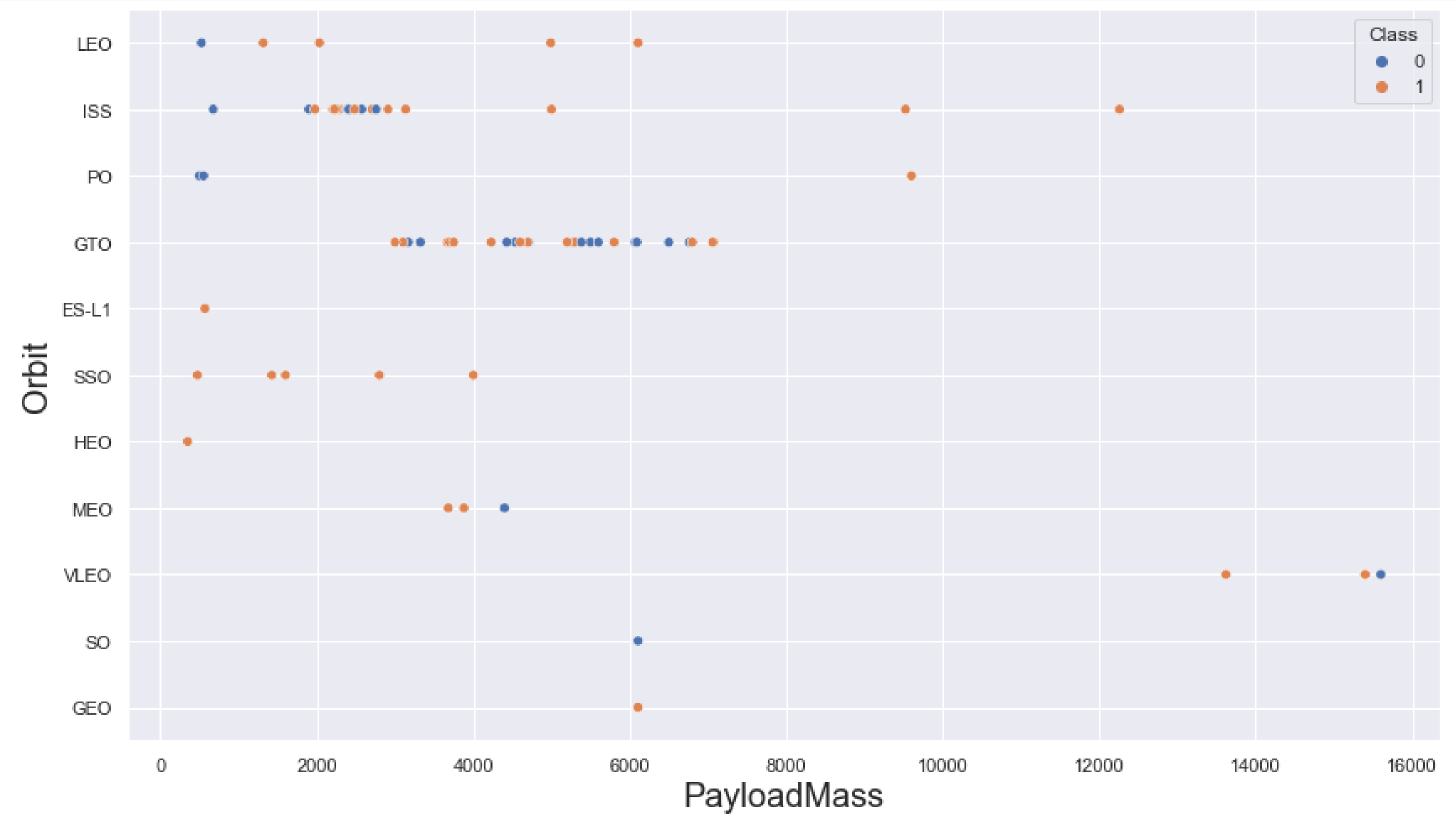
- The LEO orbit had the highest success rate with a higher number of flights.



[Github Link](#)

PAYLOAD VS. ORBIT TYPE

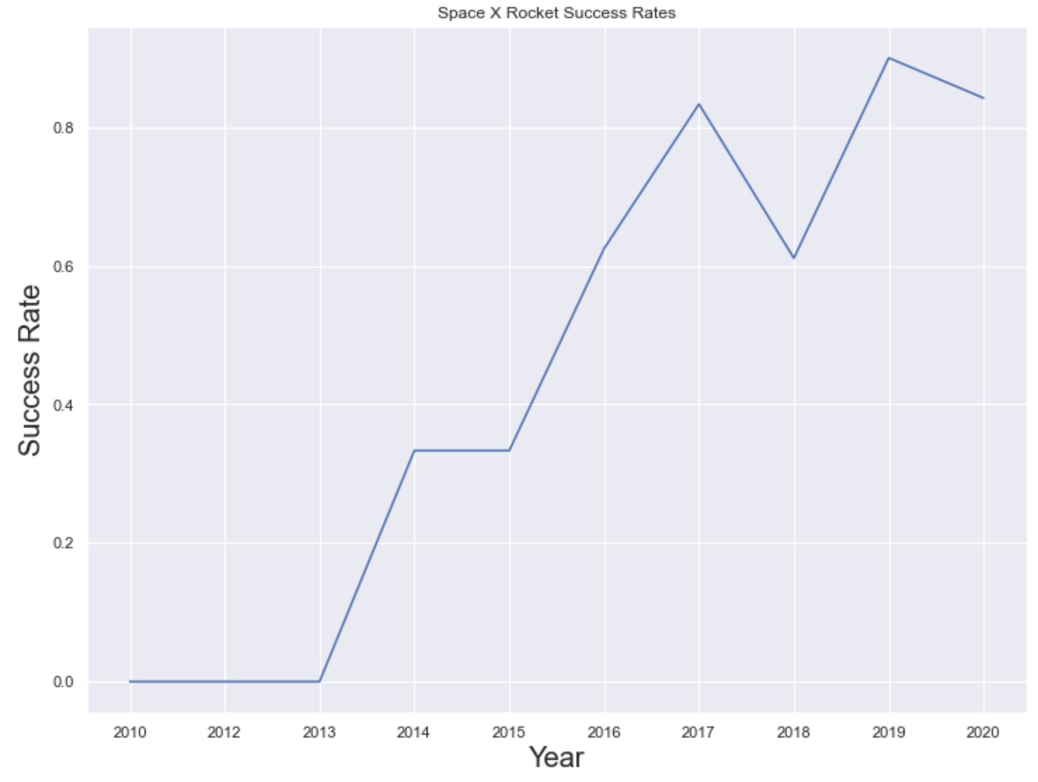
- Higher payloads negatively impact the orbits.



Github Link

LAUNCH SUCCESS YEARLY TREND

- Success rate since 2013 has increased consistently



[Github Link]

LAUNCH SITE NAMES

- Using DISTINCT in the query, we pull all the unique values for the Launch_Site column from SPACEX table.

```
In [5]: %sql SELECT DISTINCT LAUNCH_SITE as "Launch_Sites" FROM SPACEX;

 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[5]:
```

Launch_Sites
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Github Link

LAUNCH SITE NAMES WITH 'CCA'

- Using keyword 'Limit 5' in the query, we get 5 records from the SPACEX table and search with wildcard 'CCA%'.

In [6]:
```
%sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[6]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

[Github Link]

TOTAL PAYLOAD MASS

- Using the function SUM, we calculate the total in the PAYLOAD_MASS_KG_ column and WHERE clause to filter the data by the customer name.

```
In [7]: %sql SELECT SUM(PAYLOAD_MASS__KG_) AS "Total Payload Mass by NASA (CRS)" FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)';

 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
Out[7]:
```

Total Payload Mass by NASA (CRS)
45596

Github Link

AVERAGE PAYLOAD MASS BY F9 V1.1

- Using AVG, we determine the average of the column PAYLOAD_MASS_KG_ and the WHERE clause to filter the dataset.

```
In [8]: %sql SELECT AVG(PAYLOAD_MASS__KG_) AS "Average Payload Mass by Booster Version F9 v1.1" FROM SPACEX \
        WHERE BOOSTER_VERSION = 'F9 v1.1';

 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.

Out[8]:
```

Average Payload Mass by Booster Version F9 v1.1
2928

Github Link

FIRST SUCCESSFUL GROUND LANDING DATE

- Using MIN, we determine the first date in the Date column and WHERE clause to filter the data for successful landings.

In [9]:
```
%sql SELECT MIN(DATE) AS "First Succesful Landing Outcome in Ground Pad" FROM SPACEX \
WHERE LANDING__OUTCOME = 'Success (ground pad)';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[9]:

First Succesful Landing Outcome in Ground Pad
2015-12-22

Github Link

SUCCESSFUL DRONE SHIP LANDING WITH PAYLOAD (4000 – 6000)

- Selecting only Booster_Version, we use the WHERE clause to filter the dataset for successful landings and apply the payload parameters.

In [10]:
```
%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME = 'Success (drone ship)' \
AND PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[10]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Github Link

TOTAL NUMBER OF SUCCESSFUL/FAILED MISSION OUTCOMES

- Using the case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end to return the Boolean value which we get the sum.

```
In [14]: %sql SELECT sum(case when MISSION_OUTCOME LIKE '%Success%' then 1 else 0 end) AS "Successful Mission", \
             sum(case when MISSION_OUTCOME LIKE '%Failure%' then 1 else 0 end) AS "Failure Mission" \
         FROM SPACEX;

 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[14]:

Successful Mission	Failure Mission
100	1

Github Link

BOOSTERS CARRIED MAX PAYLOAD

- Using the MAX function to determine the maximum payload in the column PAYLOAD_MASS_KG_ and filter for the Booster Version

In [15]:
```
%sql SELECT DISTINCT BOOSTER_VERSION AS "Booster Versions which carried the Maximum Payload Mass" FROM SPACEX \
WHERE PAYLOAD_MASS__KG_ =(SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX);
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[15]:

Booster Versions which carried the Maximum Payload Mass
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

Github Link

2015 LAUNCH RECORDS

- List the records to display the month names, failures in drone ship, booster versions, launch_site for 2015.

In [19]:
```
%sql SELECT {fn MONTHNAME(DATE)} as "Month", BOOSTER_VERSION, LAUNCH_SITE FROM SPACEX WHERE year(DATE) = '2015' AND \
LANDING__OUTCOME = 'Failure (drone ship)';
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[19]:

Month	booster_version	launch_site
January	F9 v1.1 B1012	CCAFS LC-40
April	F9 v1.1 B1015	CCAFS LC-40

Github Link

RANK LANDING OUTCOMES BETWEEN 2010 AND 2017

- Select only LANDING_OUTCOME, WHERE clause between DATE BETWEEN 2010-06-04 and 2017-03-20. Group by Count in descending order.

In [20]:
```
%sql SELECT LANDING__OUTCOME as "Landing Outcome", COUNT(LANDING__OUTCOME) AS "Total Count" FROM SPACEX \
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' \
GROUP BY  LANDING__OUTCOME \
ORDER BY COUNT(LANDING__OUTCOME) DESC ;
```

```
 * ibm_db_sa://zpw86771:***@fbd88901-ebdb-4a4f-a32e-9822b9fb237b.c1ogj3sd0tgtu0lqde00.databases.appdomain.cloud:32731/bludb
Done.
```

Out[20]:

Landing Outcome	Total Count
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

[Github Link]

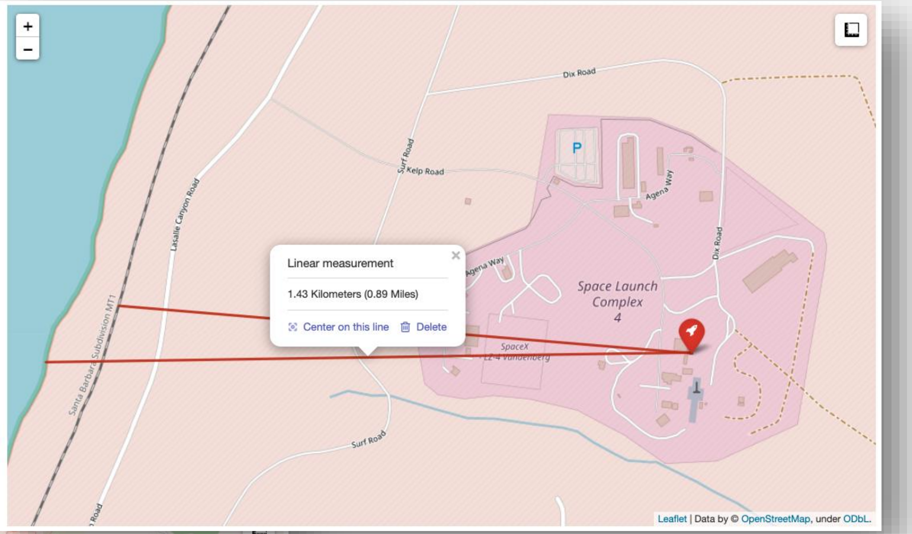
LAUNCH SITES ON FOLIUM MAP

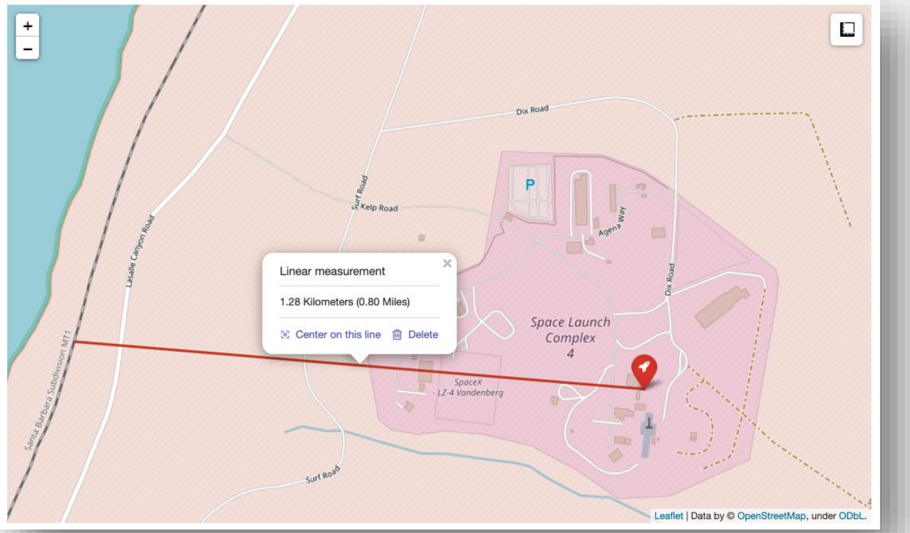
- SpaceX launch sites are near California and Florida.

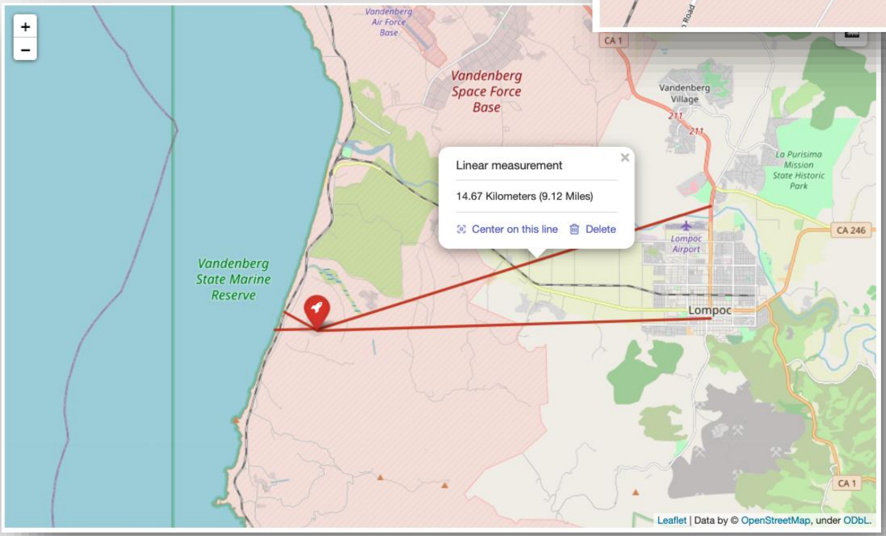
[Github Link]

LAUNCH SITE PROXIMITIES TO RAILWAYS/HIGHWAYS/COASTLINE

- Distance between all launch sites from railway tracks are greater than 1.2 KM.
- Distance between all launch sites from highways are greater than 14 KM.
- Distance between all launch sites from the coastline are greater than 1.4 KM.



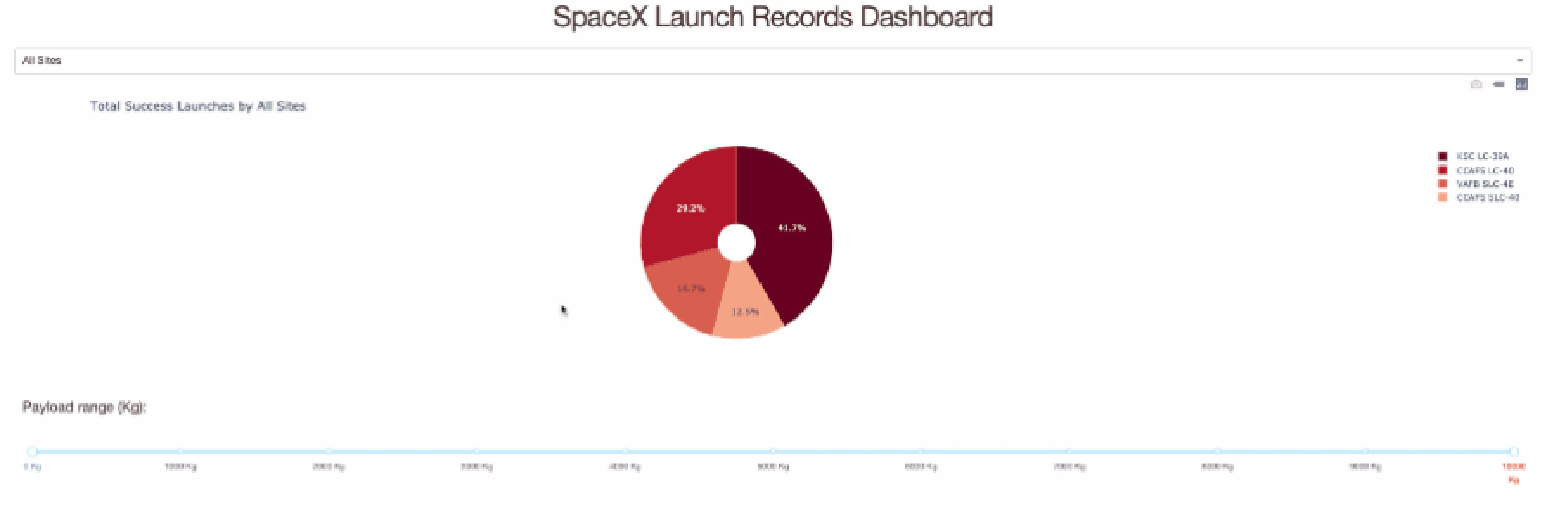




[Github Link](#)

LAUNCH SUCCESS COUNT FOR ALL SITES

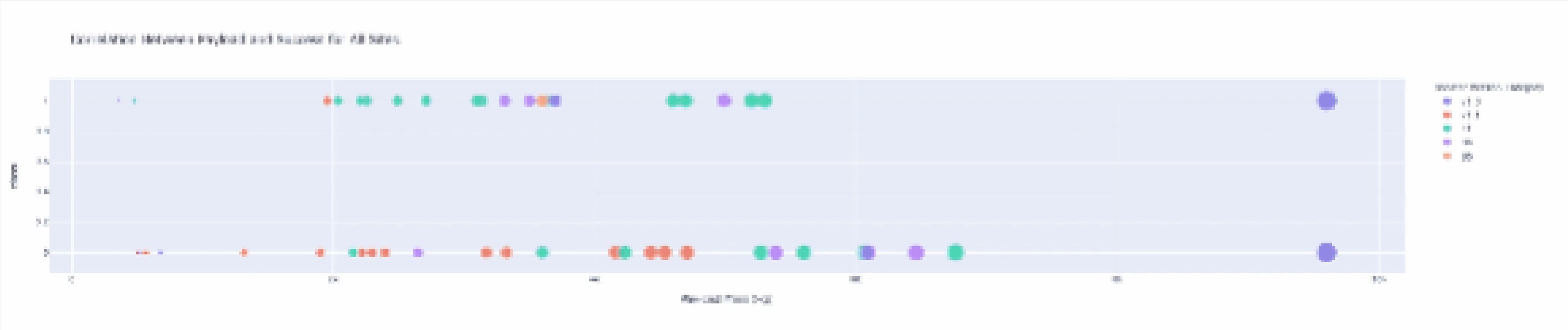
- KSC LC-39A has the most successful launches.



[Github Link]

PAYLOAD VS LAUNCH OUTCOMES

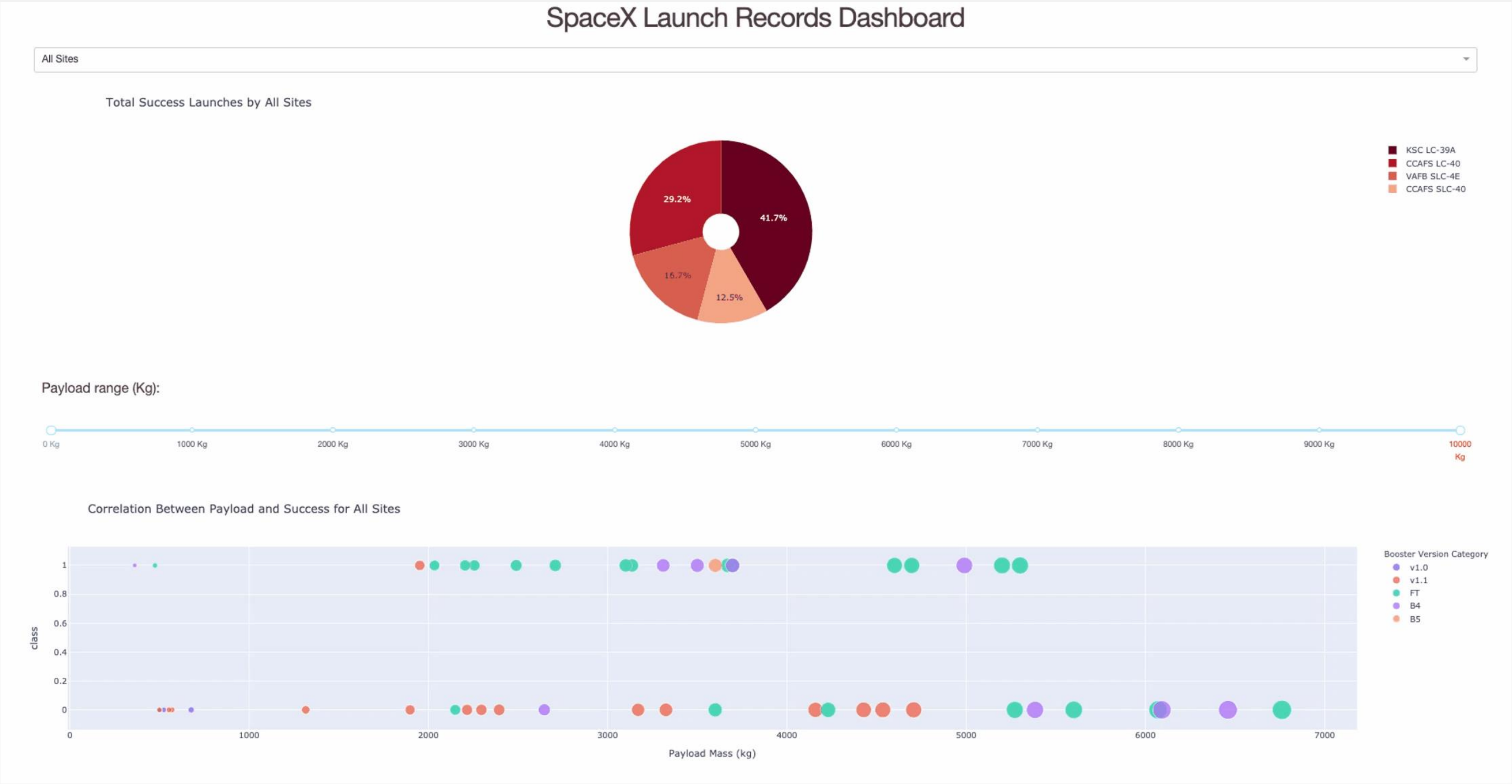
- Success rates for low weighted payloads are higher than the heavy weighted payloads.



Github Link

LAUNCH SITE WITH HIGHEST SUCCESS RATIO

- KSC LC-39A had a 76.9% success rate.
- Highest success rate:
  - Payload range 2000 KG – 10,000 KG
  - Booster version FT



Github Link

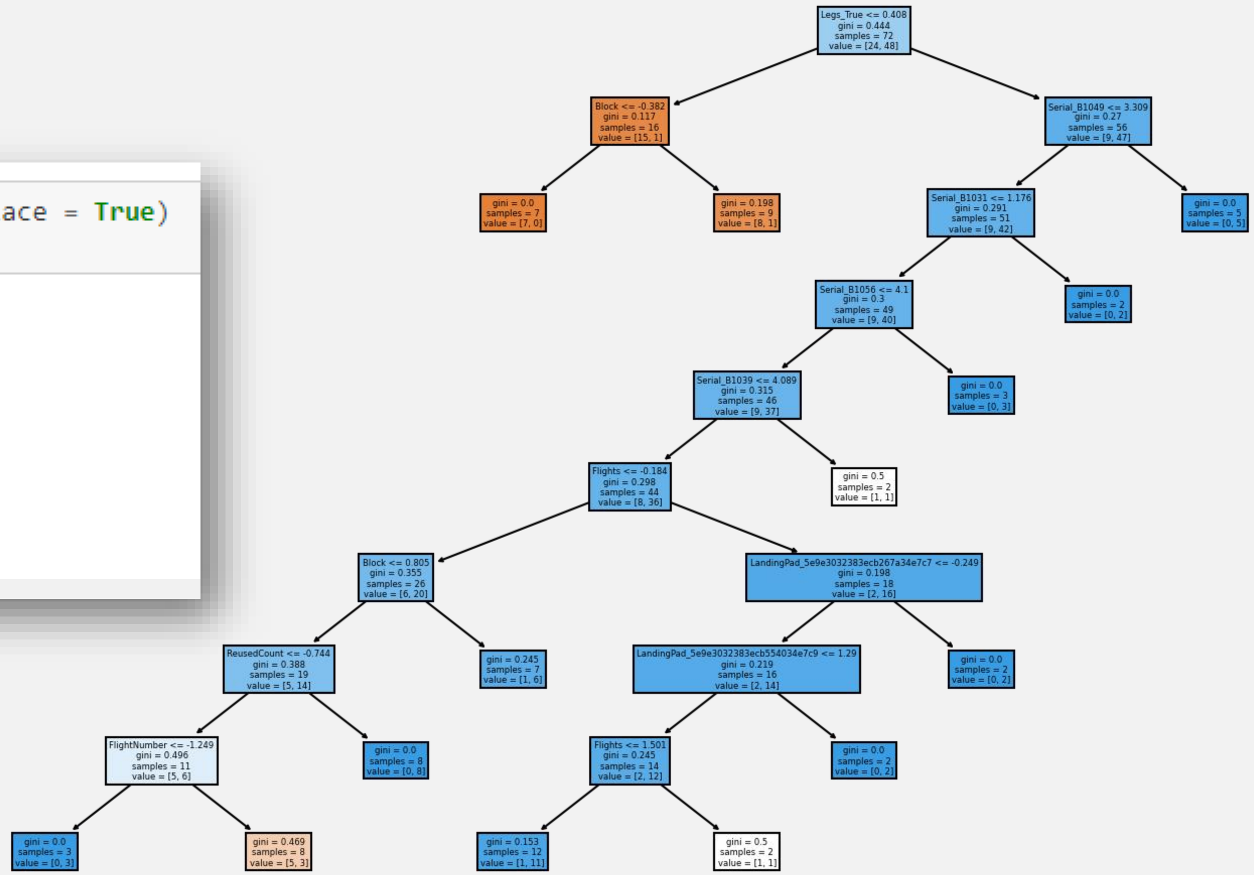
CLASSIFICATION ACCURACY

- Decision Tree gives the highest accuracy

```
In [36]: algo_df.rename(columns = {'index': 'Algorithm'}, inplace = True)
         algo_df.head()
```

Out[36]:

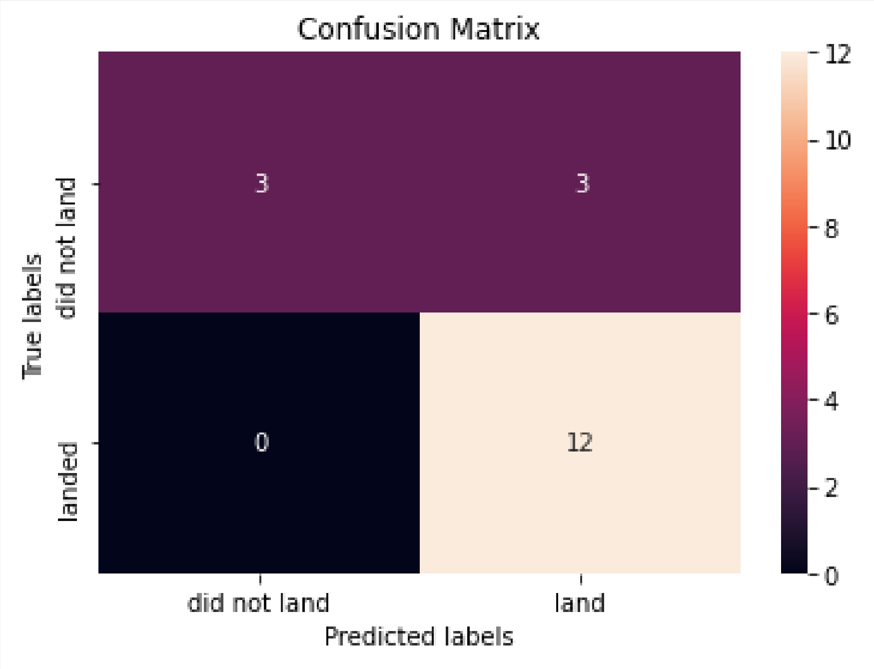
	Algorithm	Accuracy
0	Logistic Regression	0.846429
1	SVM	0.848214
2	KNN	0.848214
3	Decision Tree	0.901786



[Github Link](#)

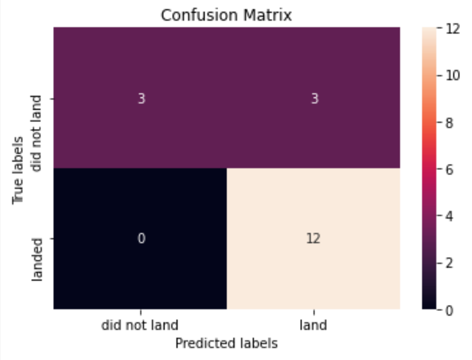
CONFUSION MATRIX

- All models have the same confusion matrix.

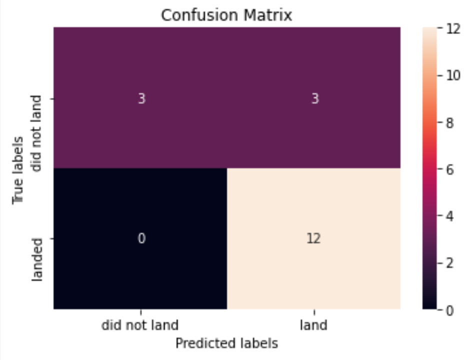


- Decision Tree

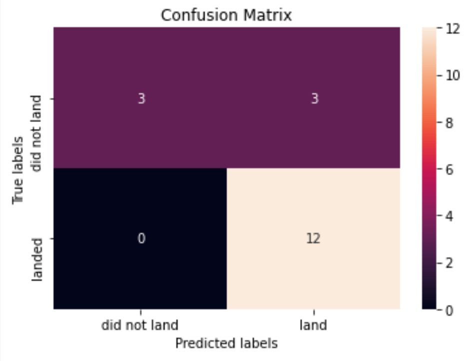
- Logistic Regression



- SVM



K Nearest Neighbor



[Github Link]

CONCLUSION

- Through this study, we have determined that Falcon9 has a strong possibility having a successful landing with a lower payload.